(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENTAMT

(12) **Patentschrift**

(10) **DE 195 10 083 C 2**

(51) Int. Cl.[6]: G 10 L 5/06

(21) Aktenzeichen: 195 10 083.2-53
(22) Anmeldetag: 20. 3. 95
(43) Offenlegungstag: 26. 9. 96
(45) Veröffentlichungstag der Patenterteilung: 24. 4. 97

Innerhalb von 3 Monaten nach Veröffentlichung der Erteilung kann Einspruch erhoben werden

(73) Patentinhaber:
International Business Machines Corp., Armonk, N.Y., US

(74) Vertreter:
Kauffmann, W., Dipl.Phys. Dr., Pat.-Ass., 70569 Stuttgart

(72) Erfinder:
Spies, Markus, Dr., 69118 Heidelberg, DE

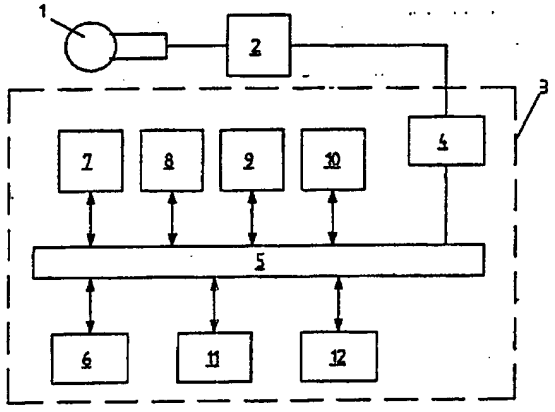
(56) Für die Beurteilung der Patentfähigkeit in Betracht gezogene Druckschriften:

DE 37 33 674 C2
WO 93 18 506

US-Z: NADAS, A.: »On Turing's Formula for World Probabilities«, In: IEEE Proc. ASSP, 33, 6, 1985, S. 1414-1416;
JELINEK, F., MERCER, R.: »Interpolated Estimation of Markov Source Parameters from Sparse Data«, In:Pattern Recognition in Practice, Amsterdam, North Holland, 1980, S. 381-397;
DE-Z: RUSKE, G.: Halbsilben als Bearbeitungseinheiten bei der automatischen Spracherkennung», In:Journal «Sprache und Datenverarbeitung», 8. Jg. 1984, Heft 1/2, S. 5-16;
SPIES, M.: «Unsicheres Wissen», Berlin,Heidelberg,1993, Spektrum Akademischer Verlag;
SPIES, Marcus: «Die Behandlung von Wortkomposita in der maschinellen Spracherkennung». In: Studien-texte zur Sprachkommunikation ISSN 0940-6832, Heft 11, S. 190-197;

(54) Verfahren und Anordnung zur Spracherkennung bei Wortkomposita enthaltenden Sprachen

(57) Verfahren zur Spracherkennung bei Wortkomposita enthaltenden Sprachen mit folgenden Schritten:
Speichern einer Menge von phonetischen Transkriptionen von Wörtern und Kompositabestandteilen;
Berechnen einer Menge von N-Gramm Häufigkeiten (Sprachmodell) für die Wahrscheinlichkeit des Auftretens eines Kompositums innerhalb einer aus N Wörtern zusammengesetzten Wort folge unter Heranziehung eines vorab verarbeiteten Textkorpus', und Speichern dieser Menge;
Erfassen und Digitalisieren des akustischen Sprachsignals sowie Speichern des digitalisierten Sprachsignals, wobei mittels einer Signalverarbeitung auf der Grundlage der phonetischen Transkriptionen näherungsweise Wörter und Kompositumbegrenzungen ermittelt werden, aus denen hypothetische Folgen von Wort- und/oder Kompositumkandidaten abgeleitet werden;
Errichten von getrennten Bearbeitungspfaden für Folgen von Kompositumkandidaten und für Folgen von Wortkandidaten;
Statistische Auswertung der Bearbeitungspfade mittels der gespeicherten N-Gramm Häufigkeiten, wobei aus der Folge der N-Gramm Häufigkeiten der Wörter bzw. Kompositabestandteile jedes Bearbeitungspfades Wahrscheinlichkeits-Profile gebildet werden; sowie
Gesamtbewertung der Bearbeitungspfade unter Heranziehung der ermittelten Wahrscheinlichkeits-Profile.



DE 195 10 083 C 2

Beschreibung

Die vorliegende Erfindung betrifft zum einen ein für Wortkomposita geeignetes Spracherkennungsverfahren, das bei sowohl diskretem als auch kontinuierlichem Diktat einsetzbar ist und das sich insbesondere zur Echtzeit-Spracherkennung eignet. Des weiteren bezieht sich die Erfindung auf eine Spracherkennungsanordnung zur Anwendung dieses Verfahrens.

Der Erfindung liegt das seitens der Anmelderin entwickelte Spracherkennungssystem TANGORA zugrunde. TANGORA ist ein Echtzeit-Spracherkennungssystem für große Vokabulare mit mehr als 20.000 Wortformen, das mit geringem Aufwand vom Benutzer sprecherspezifisch trainiert werden kann.

Ausgangspunkt bei diesem bekannten System ist die Aufteilung des Spracherkennungsprozesses in einen auf akustischen Daten basierenden Teil (Decodierung) und einen auf Sprach- bzw. Textkorpora eines bestimmten Anwendungsbereichs zurückgreifenden sprachstatistischen Teil (Sprachmodell). Die Entscheidung über Wortkandidaten ergibt sich somit jeweils aus einer Decoder- sowie einer Sprachmodell-Wahrscheinlichkeit. Für den Anwender ist primär die aufgrund dieser Architektur mögliche Anpassung des vom Erkennungssystem verarbeiteten Wortschatzes an branchenspezifische oder sogar individuelle Anforderungen von besonderer Bedeutung.

Bei diesem Spracherkennungssystem liefert die akustische Decodierung zunächst Worthypothesen. Bei der weiteren Bewertung miteinander konkurrierender Worthypothesen werden nun die Sprachmodelle zugrundegelegt. Diese stellen aus anwendungsspezifischen Textkorpora gewonnene Schätzungen von Wortfolgenhäufigkeiten dar und basieren auf einer Sammlung von Textproben aus einem gewünschten Anwendungsbereich. Aus diesen Textproben werden die häufigsten Wortformen und Wortfolgestatistiken generiert.

Bei dem hier angewandten Verfahren zur Häufigkeitsschätzung von Wortfolgen werden die Häufigkeiten für das Auftreten von sogenannten Wortform-Trigrammen in einem gegebenen Textkorpus geschätzt (siehe u. a. Nadas, A., "On Turing's Formula for Word Probabilities", IEEE Proc. ASSP, 33, 6, 1985, pp. 1414—1416). Bei einem Wortschatz von 20.000 Wortformen, wie er derzeit in dem Spracherkennungssystem TANGORA genutzt wird, wären allerdings etwa 8 Billionen Trigramme möglich. Die in der Praxis gesammelten Korpora sind also immer noch um einige Zehnerpotenzen zu klein, um überhaupt alle Trigramme auch nur beobachten zu können.

Diesem Problem des begrenzten Wortschatzes wird a.a.O. mit der Bildung sogenannter Objektklassen, die in dem Sprachkorpus mit gleicher Häufigkeit vorkommen, begegnet. Die Schätzung basiert dabei auf der Annahme einer Binomialverteilung einer Zufallsvariablen, welche allgemein die Ziehung eines Objektes aus einer Häufigkeitsklasse beschreibt.

In bekannten Spracherkennungssystemen wird für diese zu schätzenden Wahrscheinlichkeiten häufig das sogenannte Hidden-Markov-Modell angewendet. Hierbei werden mehrere im Textkorpus beobachtete Häufigkeiten zugrundegelegt. Für ein Trigramm "uvw" sind dies ein Nullgramm-Term f_0, ein Unigramm-Term f(w), ein Bigramm-Term f (w|v) sowie ein Trigramm-Term f (w|uv). Diese Terme entsprechen den im Textkorpus beobachteten relativen Häufigkeiten, wobei dem Nullgramm-Term lediglich eine korrektive Bedeutung zukommt.

Faßt man diese Terme als Wahrscheinlichkeiten des Wortes w unter verschiedenen Bedingungen auf, so kann man eine sogenannte latente Variable zufügen, von der aus durch Zustandsübergänge eine der vier Bedingungen erreicht wird, die das Wort w erzeugen. Bezeichnet man die Übergangswahrscheinlichkeiten für die betreffenden Terme mit $\lambda_0\ \lambda_1\ \lambda_2\ \lambda_3$, so ergibt sich folgender Ansatz für die Darstellung der gesuchten Trigrammwahrscheinlichkeit

$$Pr(w|uv) = \lambda_0 f_0 + \lambda_1 f(w) + \lambda_2 f(w|v) + \lambda_3 f(w|uv) \quad (1)$$

Die eigentliche Schätzung der Übergangswahrscheinlichkeiten erfolgt mittels der Methode der sogenannten "deleted estimation" (s. Jelinek, F. und Mercer, R., "Interpolated Estimation of Markov Source Parameters from Sparse Data", in Pattern Recognition in Practice, Amsterdam, North Holland, 1980, pp. 381—397). Bei diesem Verfahren werden durch Weglassung von Korpusteilmengen mehrere kleinere Textstichproben erzeugt. Für jede Stichprobe erfolgt eine Bewertung nach der oben genannten Methode, die auf den Wortfolgestatistiken beruht.

Die bekannten Spracherkennungssysteme haben den Nachteil, daß jedes Wort als eine Wortform im Wortschatz dieser Systeme auftritt. Aus diesem Grunde werden relativ hohe Anforderungen an die Speicherkapazität der Systeme gestellt. Die im allgemeinen sehr umfangreichen Wortschätze wirken sich zudem nachteilig auf die Schnelligkeit der Erkennungsverfahren aus.

In dem Aufsatz "Halbsilben als Bearbeitungseinheiten bei der automatischen Spracherkennung", G. Ruske, Journal "Sprache und Datenverarbeitung", 8. Jahrgang 1984, Heft 1/2, S. 5—16, wird zur Lösung dieses Problems vorgeschlagen, bei der automatischen Spracherkennung zur Festlegung kleinster Bearbeitungseinheiten im Bereich der akustisch-phonetischen Analyse eine Segmentierung des Wortschatzes in Halbsilben vorzunehmen. Gegenüber Systemen, denen Silben als Grundelemente zugrundeliegen und die aus diesen Grundelementen jede sprachliche Äußerung "bausteinartig" aufbauen, weist diese Vorgehensweise hinsichtlich der Speicheranforderungen etc. Vorteile auf. Denn beispielsweise im Deutschen beträgt die Zahl der verschiedenen Silben bereits etwa 5.000. Ferner werden in dem Aufsatz die Vorzüge der silbenorientierten Segmentierung auch für die höheren Bearbeitungsstufen der Spracherkennung angesprochen, wobei von relativ sicher erkannten Silben ausgehend Worthypothesen generiert werden. Auf die Umsetzung dieser Hypothese in ein Sprachmodell wird dort allerdings nicht eingegangen.

Ein besonderes Problem bei der Spracherkennung stellen die in vielen Sprachen relativ häufig auftretenden Komposita dar. Beispielsweise treten im medizinischen Bereich häufig zusammengesetzte Fachtermini auf, die nur in einigen Sprachen durch Genitivattribute ausgedrückt werden können. Bei den bekannten Spracherkennungssytemen tritt jedes Kompositum als eine eigene Wortform im Wortschatz der Systeme auf, woraus sich

Nachteile bezüglich der Performance dieser Systeme, beispielsweise aufgrund der erhöhten Anforderungen an den Speicher, ergeben.

In der internationalen Patentanmeldung WO 93/18506, DRAGON SYSTEMS INC., USA, ist ein Spracherkennungssystem für Komposita enthaltende Sprachen vorveröffentlicht, dem das vorgenannte Problem des Speicherzuwachses zugrundeliegt und das die Aufnahme von Komposita in das aktive Vokabular vermeiden will. Zur Lösung wird vorgeschlagen, eine spezielle Erkennungseinrichtung für Komposita einzusetzen. Bei einem möglicherweise vorliegenden Kompositum wechselt diese Einrichtung in einen besonderen Betriebsmodus, in dem mögliche Kompositum-Kandidaten in Form einer Liste angezeigt werden, aus der der Benutzer das richtige Kompositum manuell auszuwählen hat.

Der vorliegenden Erfindung liegt somit die bereits in dem in Studientexte zur Sprachkommunikation ISSN 0940—6832, Heft 11, S. 190—197, vorveröffentlichten Aufsatz von M. Spies mit dem Titel "Die Behandlung von Wortkomposita in der maschinellen Spracherkennung", genannte Aufgabe zugrunde, ein Verfahren bzw. eine Anordnung zur Spracherkennung bereitzustellen, bei denen vermieden wird, daß Komposita jeweils als Ganzes im Sprachmodell berücksichtigt werden müssen. Im Gegensatz dazu sollen nur Bestandteile von Komposita Berücksichtigung finden. Darüber hinaus soll eine voll maschinelle Erkennung auch von Komposita ermöglicht werden.

Diese Aufgabe wird bei dem erfindungsgemäßen Spracherkennungsverfahren gelöst durch die im Patentanspruch 1 vorgeschlagenen Verfahrensschritte.

Das erfindungsgemäße Spracherkennungsverfahren geht von dem Ansatz aus, im Sprachmodell nicht vollständige Komposita zu speichern, sondern lediglich Kompositabestandteile in Form von Einzelwörtern. Das Erkennungssystem hat demnach nur diese Bestandteile im Vokabular zu verwalten. Ein wesentlicher Gesichtspunkt dieses Lösungsgedankens ist, bei der Erkennung möglicher Komposita für die entsprechenden Kompositabestandteile sowie für die möglichen Einzelwörter getrennte Bearbeitungspfade einzurichten, d. h. eine jeweils unterschiedliche Weiterverarbeitung der hypothetischen zeitlichen Abfolgen von Wortkandidaten, die im Verlauf der Spracherkennung aus einer Folge phonetischer Transkriptionen von Wörtern und Kompositabestandteilen generiert werden. Auf diesen Bearbeitungspfaden werden dann für Komposita spezifische Sprachmodellstatistiken zur Bewertung der Worthypothesen berechnet.

Bei den N-Gramm Statistiken hat es sich als besonders vorteilhaft erwiesen, Wortform-Trigramme zu verwenden. Die Verwendung von Trigrammen im Sprachmodell hat den Vorteil, daß ein idealer Kompromiß zwischen Speicherbelastung und Verarbeitungsgeschwindigkeit geschaffen wird.

Bei dem erfindungsgemäßen Spracherkennungsverfahren können ferner für einen Kompositumendteil-Kandidaten W, gegeben einen Kontext C, im Sprachmodell distante N-Grammhäufigkeiten Pr(W/C) nicht unmittelbar benachbarter Teile einer Wortfolge gebildet werden. Grundlage dieser Sprachmodellstatistik ist eine Zerlegung der Wahrscheinlichkeiten, bei der der vorausgehende Kontext und die Bestandteile eines Kompositums getrennt berücksichtigt werden. Einen Schlüssel zur Lösung dieses Problems liefert wieder die in der Linguistik bekannte Tatsache, daß grammatisch bestimmende Teile eines Kompositums in der Regel am Kompositumende aufzufinden sind, wobei diese Bestandteile Auskunft über Genus, Casus, Numerus geben, sofern das Kompositum ein Substantiv ist. Analoges gilt jedoch auch bei aus mehreren Wörtern zusammengesetzten Verben.

Eine Verallgemeinerung dieser Tatsache führt zu der Sprachmodellannahme, daß der einem Kompositum vorausgehende Kontext die Wahrscheinlichkeit des letzten Kompositumbestandteils stark beeinflußt und daß umgekehrt, kennt man diesen letzten Bestandteil, der vorausgehende Kontext wenig über die übrigen Kompositumbestandteile aussagt. Im Sprachmodell entspricht dies einer N-Grammhäufigkeit Pr(W/C), d. h. der Wahrscheinlichkeit des letzten Bestandteils W eines Kompositums, gegeben den Kontext C. Der letzte Bestandteil W und der Kontext C sind dabei nicht unmittelbar benachbarte Teile der betrachteten Wortfolge.

Bei dem erfindungsgemäßen Spracherkennungsverfahren können ferner für einen Kompositumendteil-Kandidaten W, gegeben einen Kompositumanfang A, im Sprachmodell interne N-Grammhäufigkeiten Pr(A/W) mit inverser zeitlicher Abfolge der Kompositumbestandteile gebildet werden. Die sogenannte interne N-Grammhäufigkeit Pr(A/W) repräsentiert dabei die Häufigkeit des Kompositumanfangs A, gegeben das Kompositumende W. Die hier in umgekehrter Zeitrichtung verlaufende Wahrscheinlichkeitsannahme beruht wiederum auf der bereits genannten Tatsache, daß in den meisten Sprachen die grammatisch bestimmenden Teile eines Kompositumes regelmäßig am Kompositumende stehen.

Bei dem erfindungsgemäßen Spracherkennungsverfahren kann ferner vorgesehen sein, daß die Bewertung des Sprachkontextes sowohl auf Komposita als auch auf Kompositabestandteilen beruht. Unter der oben genannten Wahrscheinlichkeitsannahme läßt sich hiermit die Einbeziehung des Kontextes in dem der Erfindung zugrundeliegenden Sprachmodell vielseitiger gestalten. Eine Bewertung basierend auf Kompositabestandteilen bietet sich insbesondere dann an, wenn der Kontext Mehrfachkomposita enthält.

Bei dem erfindungsgemäßen Spracherkennungsverfahren kann weiter vorgesehen sein, daß akustische Verschleifungen oder Kontraktionen benachbarter Wörter mittels einer Kontextfunktion berücksichtigt werden. Bei benachbarten Wortanfängen und Wortenden, insbesondere bei Kompositaanfängen und Kompositaenden, tritt regelmäßig eine gegenseitige Beeinflussung der jeweiligen Aussprache dieser Wortteile auf. Dies rührt letztlich daher, daß in den meisten Sprachen grundsätzlich ein Bestreben festzustellen ist, bei der Aussprache benachbarter Wörter bzw. Kompositabestandteile diese möglichst übergangslos und ohne Pausen aneinanderzureihen. Dieses Problem wird aufgrund der vorgeschlagenen Kontextfunktion sehr vorteilhaft gelöst.

Bei dem erfindungsgemäßen Spracherkennungsverfahren kann ferner vorgesehen sein, daß für Kompositumkandidaten ein Bearbeitungspfad bereits dann angelegt wird, wenn ein potentieller Anfangsteil aufgrund einer spezifischen Pfadbewertung zu einer Kompositumhypothese beobachtet wird. Daher kann ein sogenanntes Likelihoodprofil unter der Hypothese, es handle sich um ein Kompositum, berechnet werden. Das Likelihoodprofil stellt ein Maß für die Qualität eines Bearbeitungspfades dar. Trifft die Kompositumhypothese zu, sollte

dieses Profil günstiger ausfallen als das alternativer Pfade. Hierdurch wird die Automatisierung des Spracherkennungsprozesses erheblich vereinfacht.

Bei dem erfindungsgemäßen Spracherkennungsverfahren kann ferner vorgesehen sein, daß das Sprachsignal mittels einer Grobabstimmung zur Ermittlung wahrscheinlicher Wort- bzw. Kompositumgrenzen ausgewertet wird, und daran anschließend eine Feinabstimmung zwischen dem akustischen Signal und den jeweiligen Wort- bzw. Kompositumkandidaten vorgenommen wird. Bei der Grobabstimmung werden Wort- bzw. Kompositumkandidaten sowie Zeitpunkte wahrscheinlicher Grenzen von Wörtern und/oder Kompositabestandteilen ermittelt und diese Ergebnisse dahingehend geprüft, ob Annäherungen an Kompositumbestandteile vorliegen und inwieweit die Kompositumkandidaten anhand der Sprachmodellwahrscheinlichkeiten mit den gegebenen Bearbeitungspfaden übereinstimmen. Bei der im Anschluß daran durchgeführten Feinabstimmung wird die Gesamtabfolge etwa ermittelter Komposita — eventuell unter Berücksichtigung von Verschleifungen anhand der Kontextfunktion — nochmals mit dem akustischen Sprachsignal verglichen und deren Übereinstimmung geprüft.

Bei dem erfindungsgemäßen Spracherkennungsverfahren kann ferner vorgesehen sein, daß für jeden Bearbeitungspfad Zugriffe auf relevante Sprachmodelldatenblöcke erfolgen. Hierdurch wird verhindert, daß bei jeder Prüfung auf einem Bearbeitungspfad ständig das vollständige Sprachmodell bereitgestellt werden muß. Aufgrund dieses Zugriffs auf Datenblöcke wird demnach die Verarbeitungsgeschwindigkeit des Erkennungssystems weiter erhöht.

Die Vorzüge der weiteren, in den Unteransprüchen 9 bis 11 charakterisierten Ausführungsbeispiele der Erfindung gegenüber dem Stand der Technik werden in der figurativen Beschreibung ausführlich erörtert.

Gegenstand der vorliegenden Erfindung ist zudem eine Spracherkennungsanordnung, bei der das erfindungsgemäße Spracherkennungsverfahrens zur Anwendung kommt. Diese Anordnung weist erfindungsgemäß eine Einrichtung zur Erfassung des akustischen Sprachsignals, eine Einrichtung zur Digitalisierung des analogen akustischen Sprachsignals, eine Einrichtung zur Erstellung einer Menge von phonetischen Transkriptionen von Wörtern und Kompositabestandteilen, eine Einrichtung zur Erstellung von Listen bezüglich einfacher Wörter, Kompositumanfangsteile und Kompositumendteile, eine Einrichtung zur Ermittlung der jeweiligen Sprachmodellwahrscheinlichkeiten auf einem Bearbeitungspfad für diese drei Listen, eine Einrichtung zur Ermittlung von Wahrscheinlichkeits-Profilen für hypothetische Folgen von Wort- und/oder Kompositionskandidaten und eine Einrichtung zur Erzeugung und Vernichtung von Bearbeitungspfaden sowie zur Entscheidung über die Erzeugung und die Vernichtung von Bearbeitungspfaden auf. Im Rahmen des Spracherkennungsprozesses wird jede Liste unter verschiedenen Bedingungen, z. B. Kontexten, geprüft.

Ein Vorteil dieser Anordnung gegenüber Spracherkennungssystemen nach dem Stand der Technik ist die vollständige Automatisierbarkeit des Spracherkennungsprozesses, unabhängig von den Diktatbedingungen. Weiterhin kann die Spracherkennung in Echtzeit erfolgen. Weitere Vorteile der Erfindung ergeben sich aus der figurativen Beschreibung.

Bei der erfindungsgemäßen Spracherkennungsanordnung kann ferner eine Einrichtung zur Kennzeichnung von Kompositabestandteilen als Anfangs- oder Schlußteile vorgesehen sein. Die Kennzeichnung kann beispielsweise in Form einer Flagge erfolgen. Ein Vorteil dieser Anordnung ist die Erhöhung der Schnelligkeit dieses Erkennungs-Teilprozesses, wodurch auch die Performance des gesamten Systems gesteigert wird.

Auf die vorteilhaften Ausgestaltungen der erfindungsgemäßen Spracherkennungsanordnung gemäß den Unteransprüchen 14 bis 16 wird im figurativen Beschreibungsteil näher eingegangen.

Das Spracherkennungsverfahren sowie die Anordnung zur Spracherkennung gemäß der Erfindung werden nachfolgend anhand von Zeichnungen am Beispiel der Kompositabehandlung in der deutschen Sprache eingehender beschrieben.

Im einzelnen zeigen:

Fig. 1 eine schematische Darstellung der erfindungsgemäßen Spracherkennungsanordnung; und

Fig. 2 die Funktionsweise der Spracherkennungsanordnung gemäß Fig. 1 bei der Erkennung von deutschsprachigen Wortkomposita anhand eines schematischen Blockdiagramms.

Bei der in Fig. 1 dargestellten Spracherkennungsanordnung wird das Sprachsignal zunächst mittels eines Mikrofons 1 erfaßt. Anstelle der Verwendung eines Mikrofons kann das Sprachsignal allerdings auch vorab auf einem Speichermedium, beispielsweise einem Diktiergerät, zwischengespeichert sein. Dieses Signal wird mittels eines Analog/Digital-Wandlers 2 in ein elektronisch weiterverarbeitbares digitales Signal umgewandelt.

Die Weiterverarbeitung des digitalen Signals erfolgt mittels einer Prozessoreinheit 3. Über einen Eingangskanal 4 gelangt das digitale Signal auf eine Sammelleitung 5 der Prozessoreinheit 3, über die eine Prozessor-Zentraleinheit 6, Speicher 7, 8, 9, 10, ein Decoder 11 und ein Likelihood-Prozessor 12 miteinander kommunizieren.

Die Speicher 7, 8, 9, 10 können jedoch auch in eine einzelne Speichereinheit integriert sein. Im Speicher 7 sind die bei der akustischen Signalverarbeitung im Decoder 11 zugrundegelegten phonetischen Transkriptionen abgelegt. Letztere stellen akustisch-phonetische Abbilder gesprochener Worte dar. Im Speicher 8 sind beispielsweise mittels der Zentraleinheit 6 vorab erstellte Listen einfacher Wörter, Kompositumsanfangs- und -endteile abgelegt. Die dem Sprachmodell zugrundeliegenden N-Gramm Häufigkeiten befinden sich im Speicher 9 und wurden vorab aus für den jeweiligen Anwendungsbereich spezifischen Textkorpora gebildet. Im Speicher 10 wird schließlich das zu untersuchende digitale Sprachsignal gespeichert.

Bei der Spracherkennung von Komposita gemäß der Blockdarstellung in Fig. 2 sei zunächst angenommen, daß die Kompositabestandteile zusammenhängend diktiert werden, wobei die Übergänge zwischen Kompositabestandteilen akustisch anders ausfallen werden, als bei einem diskreten Diktat. Mittels einer Grobabstimmung 20, die in erster Annäherung aufgrund eines vorgegebenen Vokabulars Kompositakandidaten identifiziert, werden zunächst Zeitpunkte wahrscheinlicher Wort- bzw. Kompositagrenzen ermittelt. Da die Kompositabestandteile als einzelne Wörter im Vokabular auftreten, kann die Grobabstimmung 20 am Ende eines jeden

Bestandteils einen derartigen Kompositumgrenzzeitpunkt ausmachen.

Im Anschluß daran wird anhand der Sprachmodellwahrscheinlichkeit geprüft 21, wie die bei der Grobabstimmung ermittelten Kandidaten in die gegebenen Bearbeitungspfade passen. Im Rahmen dieser Prüfung 21 kann es dann zur Anlegung von Verzweigungen 22 des Bearbeitungspfades zur Prüfung möglicher Komposita kommen. Die Verzweigung in zwei unterschiedliche Pfade stellt lediglich eine vorteilhafte Ausführungsform der Erfindung dar. Selbstverständlich sind auch Verzweigungen in drei oder mehrere Pfade denkbar.

Im weiteren wird für jeden Bearbeitungspfad 23, 24 eine Feinabstimmung 25, 26 zwischen akustischem Signal und Kompositumkandidat vorgenommen. Im Falle eines Bearbeitungspfades für eine Kompositumhypothese ("Kompositumpfad") wird dabei nach der durch die akustische Aneinanderkettung der Kompositumbestandteile des Kompositums gegebenen akustischen Symbolfolge gesucht, und nicht nach der für die einzelnen Bestandteile. Für die Berücksichtigung von Verschleifungen benachbarter Kompositumteile ist zudem eine Kontextfunktion 27 vorhanden.

Gemäß dem der Erfindung zugrundeliegenden Sprachmodell hängt die bedingte Wahrscheinlichkeit eines Kompositumbestandteils einerseits vom vorausgehenden Kontext, d. h. den dem Kompositum vorausgehenden Wörtern, andererseits von den Anfangsteilen des Kompositums selbst, ab. Die bedingte Wahrscheinlichkeit eines Kompositumanfangsteils wird dabei nicht von der desselben Wortes als Einzelwort unterschieden. Es werden lediglich je ein Bearbeitungspfad für die Einzelworthypothese sowie ein Bearbeitungspfad für die Kompositumhypothese angelegt.

Es erfolgt demnach eine Zerlegung der Wahrscheinlichkeiten, bei der der einem Kompositum vorangehende Kontext und die Bestandteile eines Kompositums getrennt berücksichtigt werden können. Ausgangspunkt für die Lösung dieses Problems liefert die von der Linguistik her bekannte Tatsache, daß im Deutschen die grammatisch bestimmenden Teile eines Kompositums regelmäßig am Kompositumende angeordnet sind. Der am Ende befindliche Bestandteil eines Kompositums gibt dabei Auskunft über Genus, Casus, Numerus, wenn das Kompositum ein Substantiv ist. Analoges gilt für Verbkomposita.

Zur Verallgemeinerung dieser Tatsache wird weiterhin angenommen, daß der vorausgehende Kontext, in dem ein Kompositum auftritt, die Wahrscheinlichkeit des letzten Bestandteils des Kompositums stark beeinflußt und daß umgekehrt, sofern der letzte Bestandteil bekannt ist, der vorausgehende Kontext wenig über die übrigen Kompositumbestandteile aussagt.

Unter der aus der Wahrscheinlichkeitstheorie abgeleiteten Annahme unabhängiger Ereignisse bedeutet dies, daß gegeben den letzten Kompositumbestandteil, die vorausgehenden Bestandteile und der vorausgehende Kontext bedingt unabhängig sind. Bezeichnet man mit W den letzten Kompositumbestandteil, mit A die vorausgehenden Bestandteile und mit C den vorausgehenden Kontext, so läßt sich eine Trigramm-Wahrscheinlichkeit des Wortes W als Kompositumendteil hinter dem Anfangsteil A im Kontext C ausdrücken als:

$$\Pr(W|CA) = \frac{\Pr(A|CW)\Pr(W|C)}{\Pr(A|C)} = \frac{\Pr(A|W)\Pr(W|C)}{\Pr(A|C)} \qquad (2)$$

In diesem Ausdruck treten zwei unterschiedliche Trigramm-Wahrscheinlichkeiten auf: Pr(A|W) und Pr(W|C), d. h. die Wahrscheinlichkeit des Kompositumanfanges A, gegeben den letzten Kompositumbestandteil W sowie die des letzten Bestandteils W, gegeben den Kontext C. Insbesondere wird hierbei ein sogenanntes distantes Trigramm (C, W) über nicht unmittelbar benachbarte Teile der Sprachäußerung gebildet. Weiterhin tritt in dem mathematischen Ausdruck (2) eine Wahrscheinlichkeit Pr(A|W) auf. Diese Wahrscheinlichkeit des Kompositumanfangsteils A, gegeben den Kompositumendteil W, entspricht demnach einer innerhalb des Kompositums durchgeführten Wahrscheinlichkeitsbetrachtung. Bemerkenswert ist hierbei, daß diese Wahrscheinlichkeiten nicht in der zeitlichen Reihenfolge der Wörter aufeinander stehen.

Bei der Implementierung dieses Sprachmodells wird ein kompositainternes Bigramm-Sprachmodell erstellt, das sogenannte Schätzer für die genannten Wahrscheinlichkeiten aus Sprachkorpora enthält, die in einem Speicher mit zugriffseffizienten Formaten abgelegt sind. Das Neuartige an diesem Modell ist, daß die kompositainternen Wahrscheinlichkeiten separat geschätzt werden, und daß diese Schätzung gegen die Zeitrichtung der gesprochenen Sprache verläuft.

Bei der technischen Ausführung des kompositainternen Modells werden drei Routinen unterschieden: Ein Zugriff auf Datenblöcke, ein Zugriff auf Daten für einzelne Kandidaten und die Berechnung der jeweiligen Pfadbewertung.

Der Zugriff auf Datenblöcke erfolgt zu Beginn der mittels eines Decoders ausgeführten akustischen Signalverarbeitung. Es liegt danach zunächst eine Reihe von Bearbeitungspfaden vor. Für jeden Pfad werden zunächst diejenigen Sprachmodelldatenblöcke gesucht, die dem vorausgehenden Kontext entsprechen. Im Falle des Kompositummodells werden, wenn ein Pfad mit einem Kompositumanfangsteil-Kandidaten endet, Datenblöcke mit den bedingten Wahrscheinlichkeiten dieses Kandidaten unter allen möglichen Schlußteilen eingelesen. Sowohl für Kompositumanfangsteile als auch für Kompositumendteile werden geeignete Flaggen eingeführt. Hiermit kann das erfindungsgemäße Spracherkennungssystem erkennen, daß ein Kompositumpfad vorliegt und für diesen Fall die entsprechenden Datenblöcke für diesen Pfad laden.

Für aktuell untersuchte Teile der Sprachäußerung wird jeweils zunächst mittels einer Grobabstimmung eine Kandidatenliste erzeugt. Dabei sind die folgenden Fälle zu unterscheiden:

1. Ist der Kompositumkandidat Anfangsteil eines potentiellen Kompositums, wird die Standard-Trigramm-

Wahrscheinlichkeit unter dem Kontext berücksichtigt. Ist diese hinreichend hoch, so wird der Bearbeitungspfad, an dessen Ende der Kompositumkandidat steht, verzweigt. Auf einem Zweig wird dann die Kompositumhypothese geprüft, auf dem anderen die des Einzelwortes.

2. Ist der Kompositumkandidat ein zweiter oder weiterer Kompositumteil eines bereits begonnen Kompositumpfades, gibt das Trigramm-Modell eine Bewertung von Null zurück. Das Kompositum-Bigramm-Modell gibt die Wahrscheinlichkeit des neuen Anfangsteils, gegeben den vorhergehenden Teil, zurück. Hierbei ist die Berechnung des sogenannten Bayesschen Theorems (Spies, M., "Unsicheres Wissen", Berlin, Heidelberg, 1993, Spektrum Akademischer Verlag) erforderlich, da die Wahrscheinlichkeiten in umgekehrter Bedingungsreihenfolge abgelegt sind.

3. Ist der Kompositumkandidat Schlußteil eines Kompositums, gibt das Trigramm-Modell die Sprachmodellwahrscheinlichkeiten des entsprechenden distanten Trigramms an. Das Kompositum-Bigramm-Modell liefert die kompositainterne Wahrscheinlichkeit des Schlußteils, gegeben dem zuletzt beobachteten Anfangsteil.

4. Kann der Kompositumkandidat sowohl Anfangsteil als auch Schlußteil sein, muß der aktuelle Bearbeitungspfad wieder verzweigt werden, zum einen für die Prüfung des Kompositumendteils, zum anderen für die des mindestens zweiten Kompositumanfangsteils. Ist diese Verzweigung vorgenommen, erfolgt für die jeweiligen Bearbeitungspfade eine Sprachmodellbewertung wie in den zuvor beschriebenen Fällen.

5. Ist der Kompositumkandidat schließlich weder Anfangs- noch Schlußteil, liefert das Kompositum-Bigrammodell eine Bewertung von Null zurück; dies führt im weiteren zum Abbruch des Kompositumpfades anhand einer Entscheidungsfunktion, die im Decoder bereits vorhanden ist.

Anwendung des Verfahrens auf Mehrfachkomposita

Es wird zunächst angenommen, daß sich der Einfluß der Kontextwörter und der Anfangsteile eines Kompositums aus den folgenden unabhängigen Teilstücken zusammensetzt:

a) Schlußteil des Kompositums, gegeben den Kontext; und
b) Anfangsteile des Kompositums, gegeben dessen Schlußteil.

Diese Zerlegung der Wahrscheinlichkeiten ist äquivalent mit der Annahme, daß bei gegebenem Kompositumschlußteil der Anfang des Kompositums unabhängig vom Kontext ist. Unter diesen Prämissen gilt mit den Anfangsteilen h1 ... hn des Kompositums, dem Schlußteil t des Kompositums, und den beiden unmittelbar dem Kompositum vorausgehenden Wörter w1 und w2, die Beziehung

$$\Pr(t|w_1 w_2 h_1 \cdots h_n) = \frac{\Pr(h_1 \cdots h_n|t)\Pr(t|w_1 w_2)}{\Pr(h_1 \cdots h_n|w_1 w_2)} \qquad (3)$$

Eine weitere Annahme besagt, daß ein Anfangsteil eines Mehrfachkompositums, der nicht zugleich Wortanfang ist, in der komposituminternen Statistik hinreichend gut durch die Wahrscheinlichkeit unter der Bedingung des unmittelbar vorausgehenden Anfangs teils beschrieben werden kann. Es gilt demnach:

$$\Pr(h_i|h_{i-1} \ldots h_1 w_1 w_2) = \Pr(h_i|h_{i-1}) \quad (n \geq i > 1) \qquad (4)$$

Schließlich wird angenommen, daß sich der Einfluß des Kompositumschlußteils auf alle Anfangsteile des Kompositums in unabhängige Beiträge des Schlußteils auf den letzten Anfangsteil und der übrigen Anfangsteile auf ihre jeweiligen Vorgänger zerlegen läßt.

$$\Pr(h_1 \ldots h_n|t) = \Pr(h_1|h_2) \ldots \Pr(h_{n-1}|h_n)\Pr(h_n|t) \qquad (5)$$

Aus diesen Annahmen läßt sich eine für die Implementierung wichtige Aussage ableiten, nämlich, daß für den ersten Kompositumanfangsteil eine Standardtrigramm-Wahrscheinlichkeit heranzuziehen ist, und daß für die darauffolgenden Anfangsteile die Wahrscheinlichkeit sich aus dem Produkt einzelner komposituminterner Bigramm-Wahrscheinlichkeiten zusammensetzt. Die entsprechende mathematische Beziehung lautet:

$$\Pr(w_1 w_2 h_1 \cdots h_n t) = \Pr(w_1)\Pr(w_2|w_1)\Pr(h_1|w_1 w_2)$$
$$\prod_{i=2}^{n} \Pr(h_i|h_{i-1}) \frac{\Pr(h_1 \cdots h_n|t)\Pr(t|w_1 w_2)}{\Pr(h_1 \cdots h_n|w_1 w_2)} \qquad (6)$$

Bei der Bearbeitung eines Mehrfachkompositums ist demnach für jeden Bestandteil jeweils nur eine vergleichsweise leicht auf suchbare Wahrscheinlichkeit in Betracht zu ziehen. Mit diesem Verfahren lassen sich

somit die Bearbeitungspfade jedes Kompositumbestandteiles korrekt bewerten.
Unter den vorgenannten Annahmen folgt schließlich die Beziehung:

$$\Pr(h_1 \cdots h_n | w_1 w_2) \doteq \Pr(h_1 | w_1 w_2) \prod_{i=2}^{n} \Pr(h_i | h_{i-1}) \qquad (7)$$

Zur Berechnung der normierten Wahrscheinlichkeit des Kompositumschlußteils müssen demnach nur die auf dem Bearbeitungspfad durch das gesamte Kompositum auftretenden Koeffizienten Pr(hi|hi−1) multipliziert werden, wodurch die Implementierung dieses Verfahrens erheblich vereinfacht wird.

Patentansprüche

1. Verfahren zur Spracherkennung bei Wortkomposita enthaltenden Sprachen mit folgenden Schritten:
Speichern einer Menge von phonetischen Transkriptionen von Wörtern und Kompositabestandteilen;
Berechnen einer Menge von N-Gramm Häufigkeiten (Sprachmodell) für die Wahrscheinlichkeit des Auftretens eines Kompositums innerhalb einer aus N Wörtern zusammengesetzten Wort folge unter Heranziehung eines vorab verarbeiteten Textkorpus', und Speichern dieser Menge;
Erfassen und Digitalisieren des akustischen Sprachsignals sowie Speichern des digitalisierten Sprachsignals, wobei mittels einer Signalverarbeitung auf der Grundlage der phonetischen Transkriptionen näherungsweise Wörter und Kompositumbegrenzungen ermittelt werden, aus denen hypothetische Folgen von Wort- und/oder Kompositumkandidaten abgeleitet werden;
Errichten von getrennten Bearbeitungspfaden für Folgen von Kompositumkandidaten und für Folgen von Wortkandidaten;
Statistische Auswertung der Bearbeitungspfade mittels der gespeicherten N-Gramm Häufigkeiten, wobei aus der Folge der N-Gramm Häufigkeiten der Wörter bzw. Kompositabestandteile jedes Bearbeitungspfades Wahrscheinlichkeits-Profile gebildet werden; sowie
Gesamtbewertung der Bearbeitungspfade unter Heranziehung der ermittelten Wahrscheinlichkeits-Profile.
2. Spracherkennungsverfahren nach Anspruch 1, dadurch gekennzeichnet, daß für einen Kompositumendteil-Kandidaten W, gegeben einen Kontext C, im Sprachmodell distante N-Grammhäufigkeiten Pr(W/C) nicht unmittelbar benachbarter Teile einer Wortfolge gebildet werden.
3. Spracherkennungsverfahren nach Anspruch 1 und/oder 2, dadurch gekennzeichnet, daß für einen Kompositumendteil-Kandidaten W, gegeben einen Kompositumanfang A, im Sprachmodell interne N-Grammhäufigkeiten Pr(A/W) mit inverser zeitlicher Abfolge der Kompositumbestandteile gebildet werden.
4. Spracherkennungsverfahren nach einem oder mehreren der Ansprüche 1 bis 3, dadurch gekennzeichnet, daß die Bewertung des Sprachkontextes sowohl auf Komposita als auch auf Kompositabestandteilen beruht.
5. Spracherkennungsverfahren nach einem oder mehreren der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß akustische Verschleifungen oder Kontraktionen benachbarter Wörter mittels einer Kontextfunktion berücksichtigt werden.
6. Spracherkennungsverfahren nach einem oder mehreren der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß für Kompositumkandidaten ein Bearbeitungspfad bereits dann angelegt wird, wenn ein potentieller Anfangsteil aufgrund einer spezifischen Pfadbewertung zu einer Kompositumhypothese beobachtet wird.
7. Spracherkennungsverfahren nach einem oder mehreren der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß das Sprachsignal mittels einer Grobabstimmung zur Ermittlung wahrscheinlicher Wort- bzw. Kompositumgrenzen ausgewertet wird, und daran anschließend eine Feinabstimmung zwischen dem akustischen Signal und den jeweiligen Wort- bzw. Kompositumkandidaten vorgenommen wird.
8. Spracherkennungsverfahren nach einem oder mehreren der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß für jeden Bearbeitungspfad Zugriffe auf relevante Sprachmodelldatenblöcke erfolgen.
9. Spracherkennungsverfahren nach einem oder mehreren der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß zur Berechnung der Wahrscheinlichkeit eines Kompositumendteils von dem vorausgehenden Kontext und dem Anfangsteil des Kompositums ausgegangen wird.
10. Spracherkennungsverfahren nach Anspruch 9, dadurch gekennzeichnet, daß eine Wahrscheinlichkeit Pr(W/CA) eines Kompositumbestandteils W als Kompositumendteil hinter einem Kompositumanfangsteil A unter Berücksichtigung des vorausgehenden, aus zwei Wörtern bzw. Komposita zusammengesetzten Kontextes C, aus dem normierten Produkt einer innerhalb des Kompositums gebildeten inneren Bigrammwahrscheinlichkeit Pr(A/W) und einer außerhalb des Kompositums gebildeten distanten Trigrammwahrscheinlichkeit Pr (W/C) ermittelt wird.
11. Spracherkennungsverfahren nach Anspruch 9 und/oder 10, bei Mehrfachkomposita enthaltenden Sprachen, dadurch gekennzeichnet, daß unter den Annahmen, daß bei gegebenem Schlußteil der Anfang eines Kompositums unabhängig vom Kontext ist, daß ein nicht am Kompositumanfang stehender Anfangsteil eines Mehrfachkompositums durch die Wahrscheinlichkeit $Pr(A_i/A_{i-1})$ seiner Folge auf den unmittelbar vorausgehenden Anfangsteil bestimmt ist, und daß sich der Einfluß des Schlußteils auf alle Anfangsteile des Kompositums in unabhängige Beiträge des Schlußteils auf den letzten Anfangsteil und der übrigen Anfangsteile auf ihre jeweiligen Vorgänger zerlegen läßt, zur Berechnung der normierten Wahrscheinlichkeit des Kompositumschlußteils auf einem Bearbeitungspfad durch das Kompositum auftretende Pfadkoeffi-

zienten multipliziert werden.

12. Anordnung zur Spracherkennung bei Wortkomposita enthaltenden Sprachen mittels eines Spracherkennungsverfahrens gemäß einem oder mehreren der vorhergehenden Ansprüche, mit
einer Einrichtung (1) zur Erfassung des akustischen Sprachsignals;
einer Einrichtung (2) zur Digitalisierung des akustischen Sprachsignals;
einer Einrichtung zur Erstellung einer Menge von phonetischen Transkriptionen von Wörtern und Kompositabestandteilen;
einer Einrichtung (6) zur Erstellung von Listen bezüglich einfacher Wörter, Kompositumanfangsteile und Kompositumendteile;
einer Einrichtung (12) zur Ermittlung der jeweiligen Sprachmodellwahrscheinlichkeiten auf einem Bearbeitungspfad für diese drei Listen;
einer Einrichtung zur Ermittlung (21) von Wahrscheinlichkeits-Profilen für hypothetische Folgen von Wort- und/oder Kompositionskandidaten; und
einer Einrichtung zur Erzeugung und Vernichtung von Bearbeitungspfaden (22) sowie zur Entscheidung über die Erzeugung und die Vernichtung von Bearbeitungspfaden.

13. Spracherkennungsanordnung nach Anspruch 12, mit einer Einrichtung zur Kennzeichnung von Kompositabestandteilen als Anfangs- oder Schlußteile.

14. Spracherkennungsanordnung nach Anspruch 12 und/oder 13, mit einer Einrichtung zum Erstellen und Laden von Datenblöcken von Sprachmodellwahrscheinlichkeiten.

15. Spracherkennungsanordnung nach einem oder mehreren der Ansprüche 12 bis 14, mit einer Einrichtung zur Bereitstellung beliebig vieler Kompositamodelle in Form von Sprachmodellklassen.

16. Spracherkennungsanordnung nach einem oder mehreren der Ansprüche 12 bis 15, mit einer Einrichtung zur Erstellung einer Kontextfunktion.

Hierzu 1 Seite(n) Zeichnungen

– Leerseite –

Nummer: DE 195 10 083 C2
Int. Cl.[6]: G 10 L 5/06
Veröffentlichungstag: 24. April 1997

FIG.1

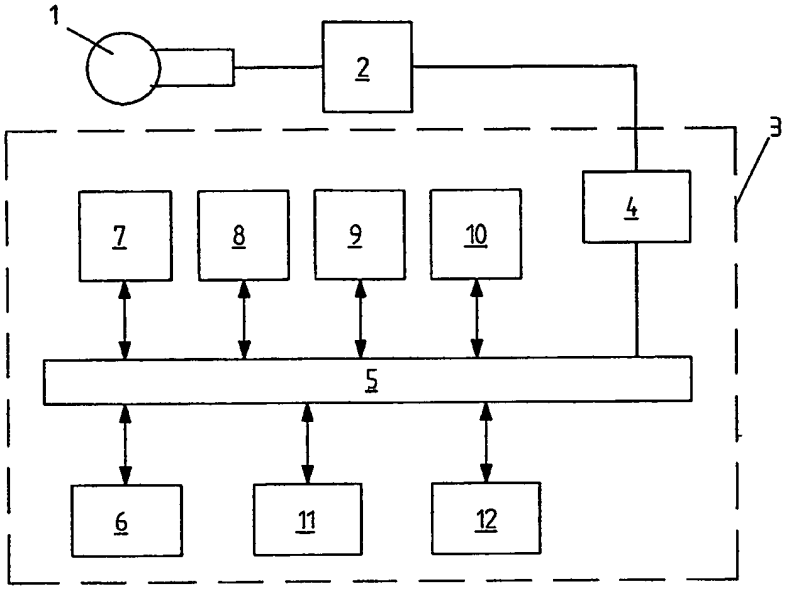


FIG.2